Bayesian Statistical Analysis and Frequentist Analysis in top statistics

Reinhard Schwienhorst

MICHIGAN STATE UNIVERSITY

January 8, 2010

Outline

- Introduction: Frequentist vs Bayesian
- Bayesian analysis
  - Posterior
  - Systematic uncertainties
- Frequentist analysis
  - Ensemble testing
- top_statistics
- Final comments

Introduction

- Physics experiments are usually out to
  - Discover something
    - Find *events* that cannot be explained by the standard model
    - Find a few events above a background
      - Statistics of small numbers
  - Measure something very precisely
    - Analyze many events in detail
    - Have very good control over the experiment
    - Systematic uncertainty

Typical Problem

- Search for events generated in some process
- The number of predicted events is given by

$$n_{pred} = acc \times lumi \times XS + n_{bkg}$$

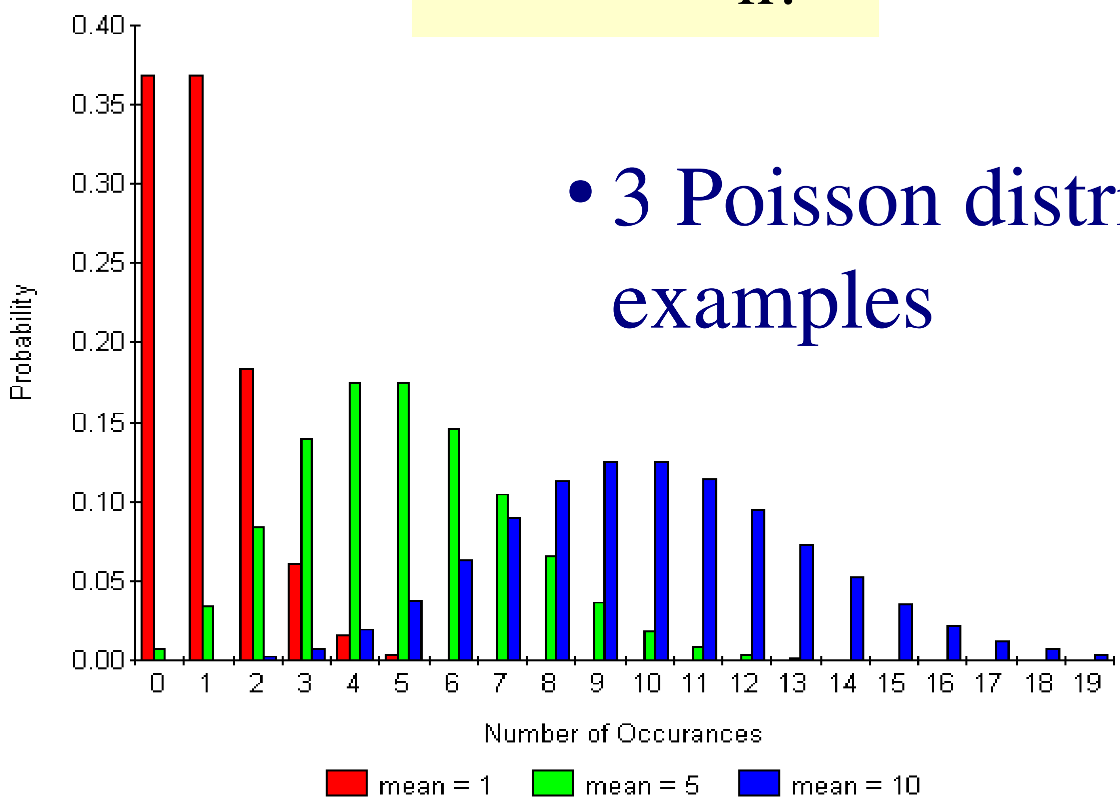
  where:
  - acc: signal acceptance, fixed and known
  - lumi: integrated luminosity, fixed and known
  - n_{bkg} : the number of background events due to ordinary SM processes, fixed and known
- The experiment tries to determine the cross section XS by relating n_{pred} to the observed events n_{obs}
- Usually either a measurement ± 1 sigma or a 90% confidence interval is given

Probability everyone can agree on

- Given a known predicted yield $\mu = n_{pred}$, what is the probability to observe count $n = n_{obs}$ in data?
- Poisson statistics

$$P(n,\mu) = \frac{\mu^n e^{-n}}{n!}$$

- 3 Poisson distribution examples



But what if I don't know the cross section and cannot predict the yield but want to determine it from the observed count?

Frequentist vs Bayesian statistics

Statistics Philosophies

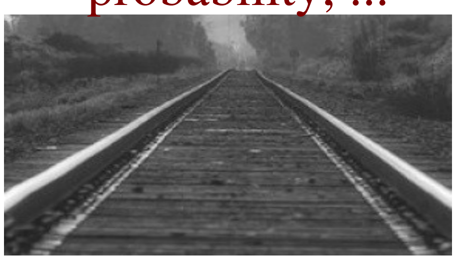
Probability is:

Frequentist

- The limiting relative frequency of a certain outcome:

$$P(A) = \lim_{n\to\infty} \frac{\text{\# of outcome A in } n \text{ measurements}}{n}$$

- True values can never be determined precisely
- Includes several assumptions
  - Experiment is repeatable, parameters don't change, each measurement has the same probability, ...



Bayesian

- Subjective:

$$P(A) = \text{degree of belief that hypothesis A is true}$$

- Intuitive definition
- Degree of belief in a measurement
- Depends on degree of belief in underlying theory

What is a 90% confidence interval?

Frequentist

- *If I repeat an experiment many times (and create a confidence interval in each experiment), the true value μ_t will lie inside the interval 90% of the time.*
- Statement about many (hypothetical) experiments
- We (Physicists) like to argue in Frequentist terms

We try to convince others in Frequentist language

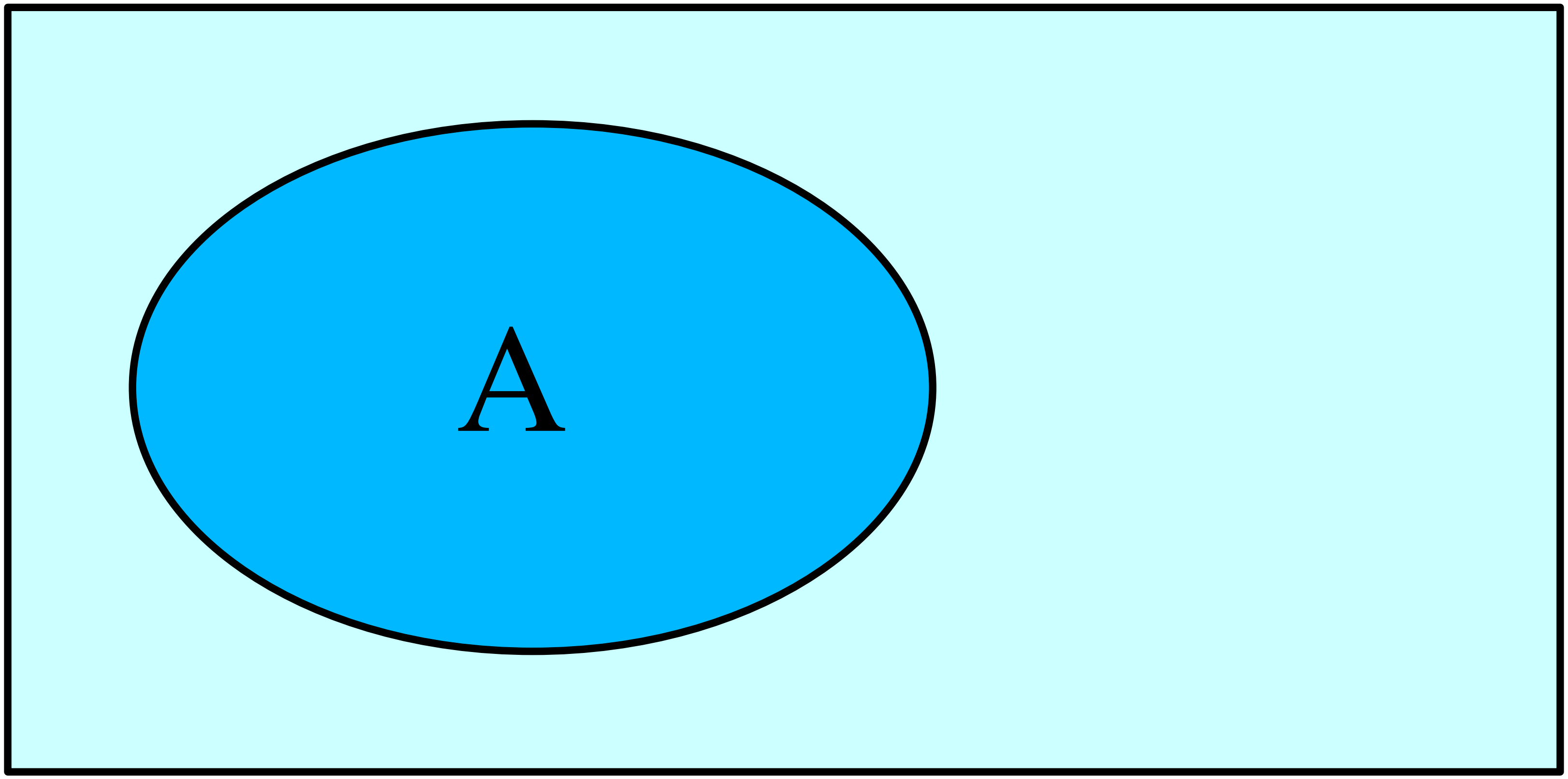
Bayesian

- *If I determine a 90% confidence level interval in a single experiment, 90% of the possible values for the true value μ_t lie inside the Bayesian interval.*
- Statement about the true value
- We (Physicists) like to think and feel in Bayesian terms

We form our own opinion with Bayesian intuition

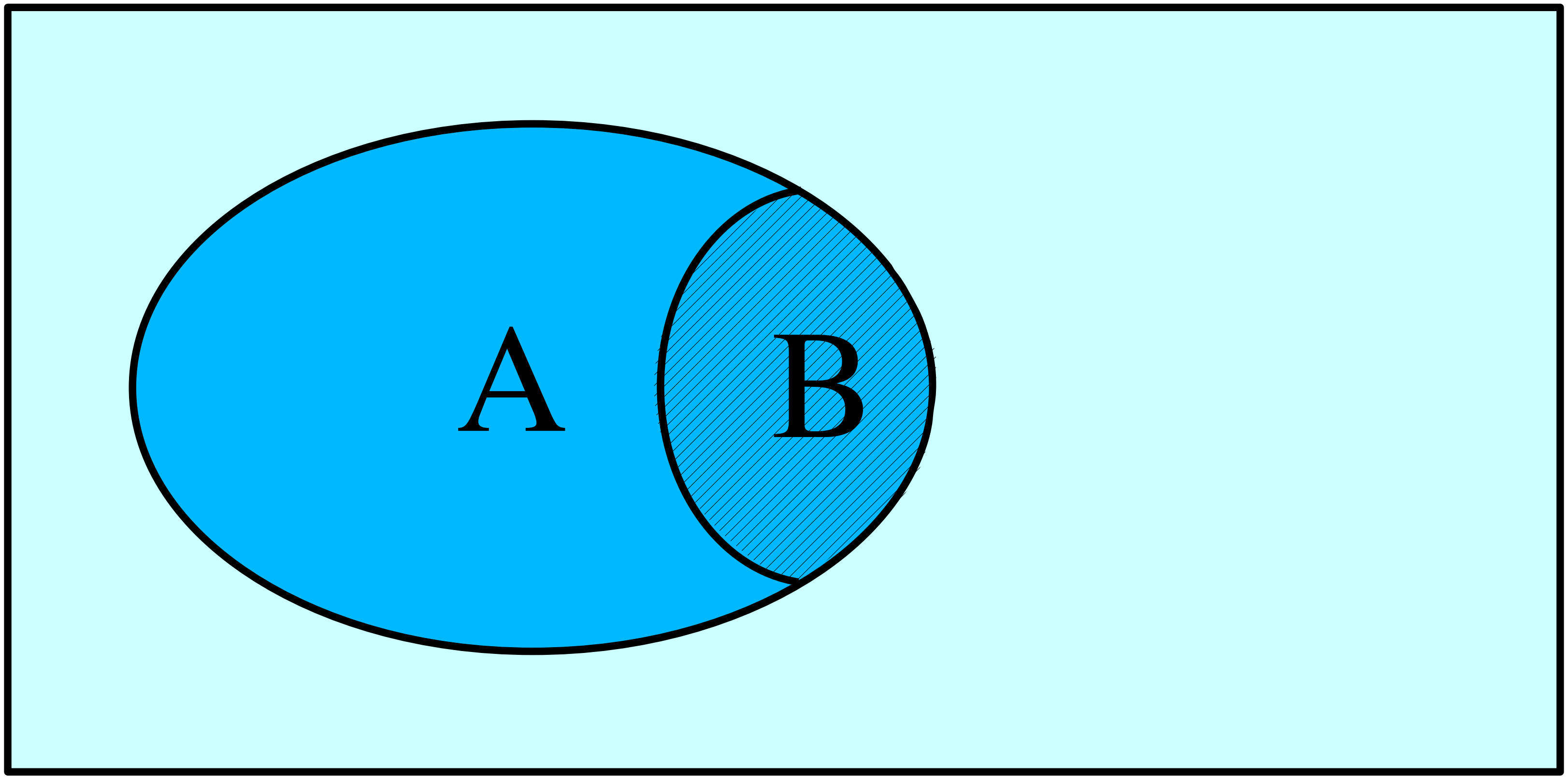
Bayesian Analysis

Simple probability



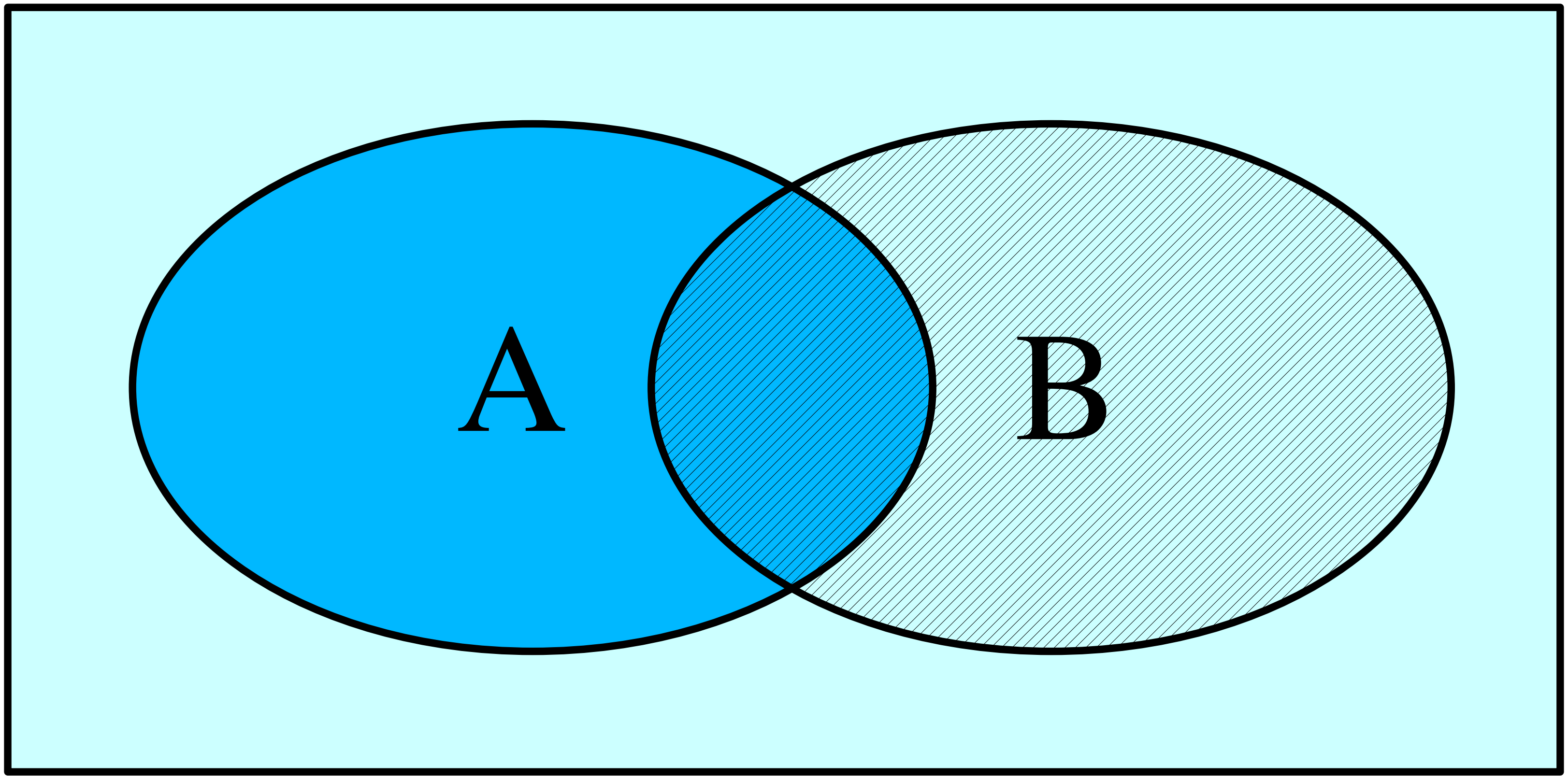
P(A): Probability that A is true

Conditional probability



P(B|A): conditional probability for B, given that A is true.

Bayes Theorem



$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$$

Bayesian Statistical Analysis

$$P(n_{pred} \mid n_{obs}) = \frac{P(n_{obs} \mid n_{pred}) \times P(n_{pred})}{P(n_{obs})}$$

- For us: n_{pred} = bkg sum + acc × lumi × XS
- If signal and data are distributed over multiple channels, take product of likelihoods in all channels

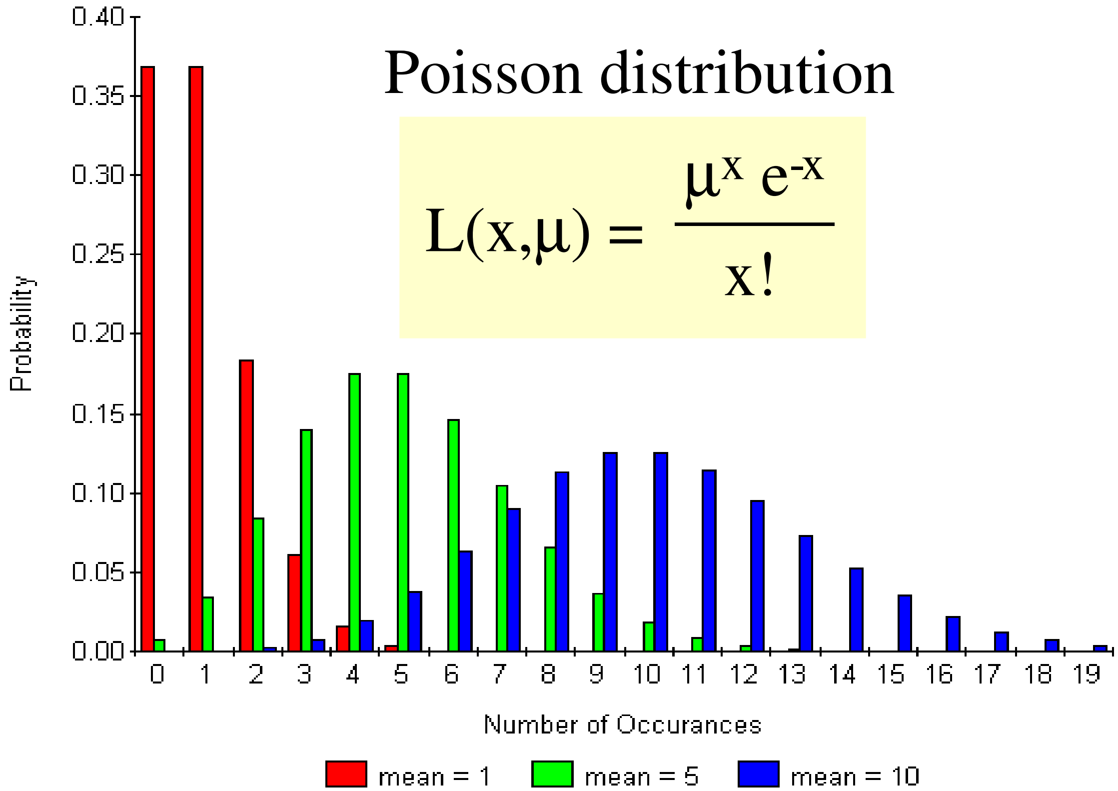
$$P_{tot} = \prod P\left(n^i_{pred} \mid n^i_{obs}\right)$$

Bayesian Statistical Analysis

$$P(n_{pred} \mid n_{obs}) = \frac{P(n_{obs} \mid n_{pred}) \times P(n_{pred})}{P(n_{obs})}$$

"Posterior probability"

Likelihood



Poisson distribution

$$L(x,\mu) = \frac{\mu^{x}\, e^{-x}}{x!}$$

Bayesian Statistical Analysis

$$P(n_{pred} \mid n_{obs}) = \frac{P(n_{obs} \mid n_{pred}) \times P(n_{pred})}{P(n_{obs})}$$

"Posterior probability" Likelihood Normalization factor "prior probability"

- Much discussion about the prior in statistics
  - Often choice is not clear
    - For example Vtb is proportional to sqrt(XS), different result if prior is flat in Vtb or in XS
  - Usually goal is "uninformed prior"
  - For us the choice is always prior flat in cross section

Bayesian Statistical Analysis

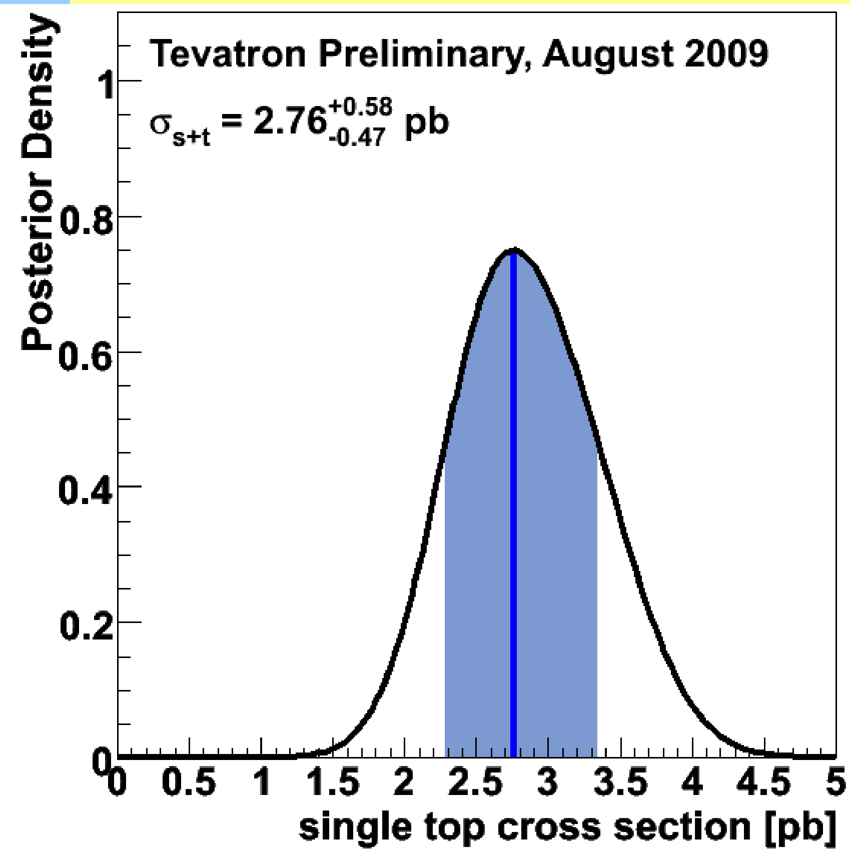
$$P(n_{pred} \mid n_{obs}) = \frac{P(n_{obs} \mid n_{pred}) \times P(n_{pred})}{P(n_{obs})}$$

"Posterior probability" Likelihood Normalization factor "prior probability"

Bayesian Statistical Analysis

$$P(n_{pred} \mid n_{obs}) = \frac{P(n_{obs} \mid n_{pred}) \times P(n_{pred})}{P(n_{obs})}$$

"Posterior probability"

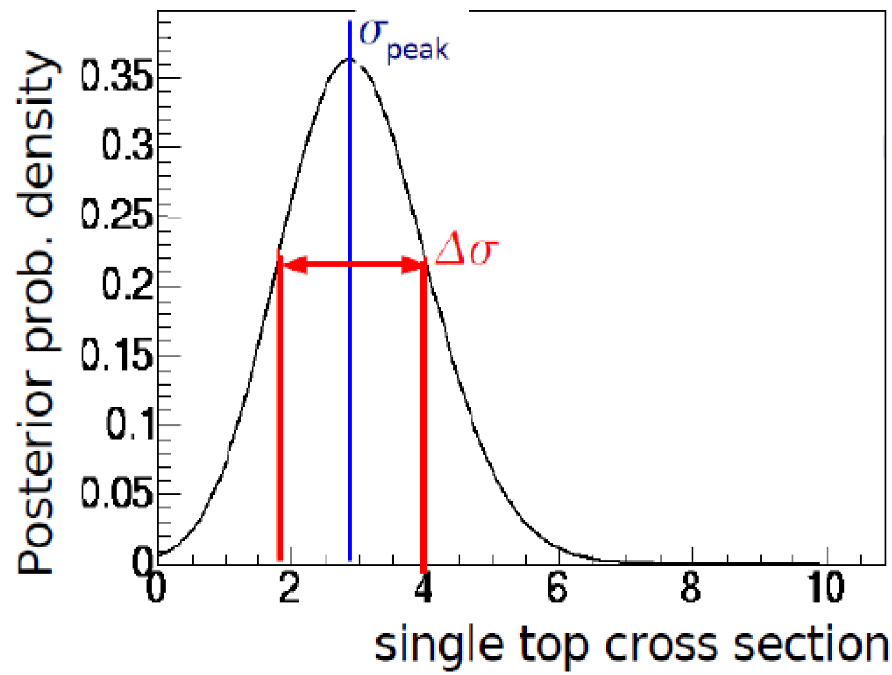


Bayesian Statistical Analysis

$$P(n_{pred} \mid n_{obs}) = \frac{P(n_{obs} \mid n_{pred}) \times P(n_{pred})}{P(n_{obs})}$$

"Posterior probability" | Likelihood | Normalization factor | "prior probability"

- Cross section
  (posterior peak)
- Cross section uncertainty
  (68% error band)
- 90% confidence level limit
  (integral from left)



Simple Bayesian example

$$\mu = n_{pred} = acc \times lumi \times XS + n_{bkg}$$

- $N_{obs} = 10$, $n_{bkg} = 7.5$, $acc \times lumi = 0.5/pb$
  - i.e. naively expect cross section of 5pb
- Compute Bayesian posterior for XS using simple spreadsheet

$$P(Nobs,\mu) = \frac{\mu^{Nobs}\ e^{-Nobs}}{Nobs!}$$

- Prior for XS is flat in XS
- Neglect posterior normalization

Simple Bayesian example

Nobs=10 Nbkg=7.5 acc*lumi=0.5/pb

XS [pb]	μ	P(Nobs|μ)
0	7.5	0.09

$$\mu = n_{pred} = acc \times lumi \times XS + n_{bkg}$$

Simple Bayesian example

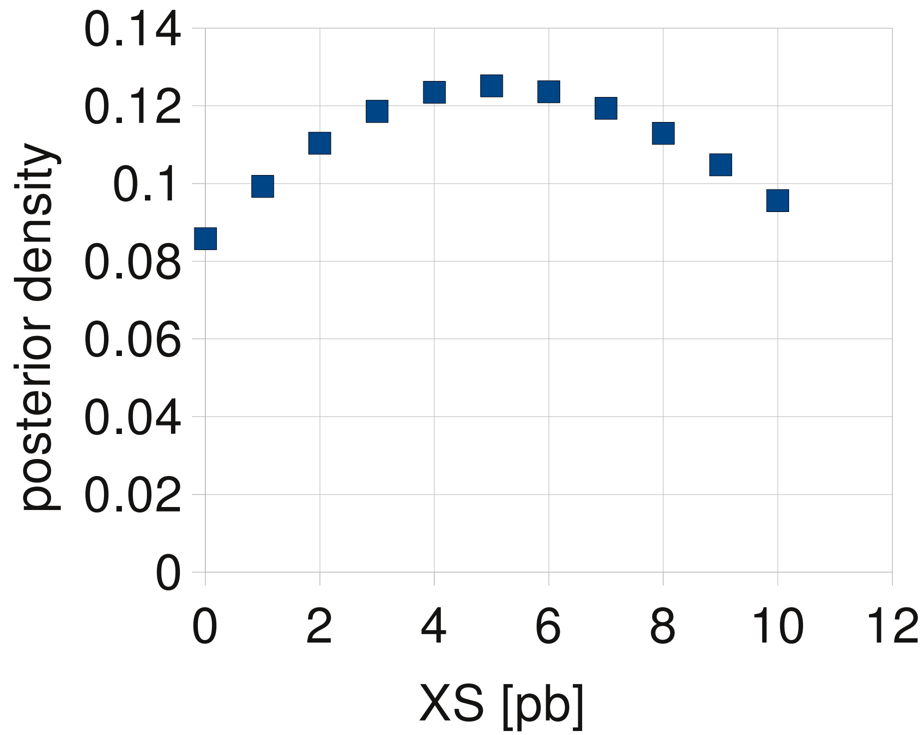
Nobs=10 Nbkg=7.5 acc*lumi=0.5/pb

XS [pb]	μ	P(Nobs|μ)
0	7.5	0.09
1	8	0.1
2	8.5	0.11
3	9	0.12
4	9.5	0.12
5	10	0.13
6	10.5	0.12
7	11	0.12
8	11.5	0.11
9	12	0.1
10	12.5	0.1

Simple Bayesian example

Nobs=10 Nbkg=7.5 acc*lumi=0.5/pb

XS [pb]	μ	P(Nobs|μ)
0	7.5	0.09
1	8	0.1
2	8.5	0.11
3	9	0.12
4	9.5	0.12
5	10	0.13
6	10.5	0.12
7	11	0.12
8	11.5	0.11
9	12	0.1
10	12.5	0.1



Simple Bayesian example in top statistics

- In climit.cpp, likelihood_generic:

```
long double y=1;
for(int ichannel = nChannels-1; ichannel >= 0; --ichannel) {
    double m = nobs[ichannel]; // Observed count for ichannel
    double s = bkg[ichannel];    // Sum for total yield in any bin
    // Add signal
    s += accL[ichannel]*x;  // x = cross-section
    // evaluate the poisson
    long double val = poisson(m, s);
    // Compute product over bins
    if(val>=0.) y *= val;
 }
return y;
```

- Multiple channel: likelihood is product over all channel
- Plus checks for invalid input numbers, y getting smaller than long double limit, etc.

Including systematic uncertainties

- Including systematics: Integrate over systematics

$$P(n_{pred} \mid n_{obs}) = \iint_{sys} \frac{P(n_{obs} \mid n_{pred}, sys) \times P(XS) \times P(sys)}{P(n_{obs})}$$

- P(sys) is a Gaussian
- Systematics either global or per channel
- Protect against "crazy" systematics
  - That are far above nominal or go below 0 yield (truncate)
- Integration using Monte Carlo sampling
  - anywhere from 2k NSamples to 1M NSamples
  - Re-draw iSample if too many bins go to 0

Systematic uncertainty integration

- Generate systematic shifts in limit_base:

```
for(NSamples systematics samples) {
  for(systematics names) {
    val = myrandom.Gaus();  // random shift for each systematic name
    sysshift[sys name] = val;
  }
}
```

- Fill background sum for each systematic sample in input.cpp, input::AddSysShiftedValue():

```
for(bins) {
   for(systematics names) {
     diff = shift*(_syst[sysname].getValuesPlus()[ibin]-value0);
     if(shift<0.) diff = shift*(value0-_syst[sysname].getValuesMinus()[ibin]);
   }
   bin_value += diff;
}
```

- Plus lognormal distribution, many checks of inputs and outputs

Systematic uncertainty integration

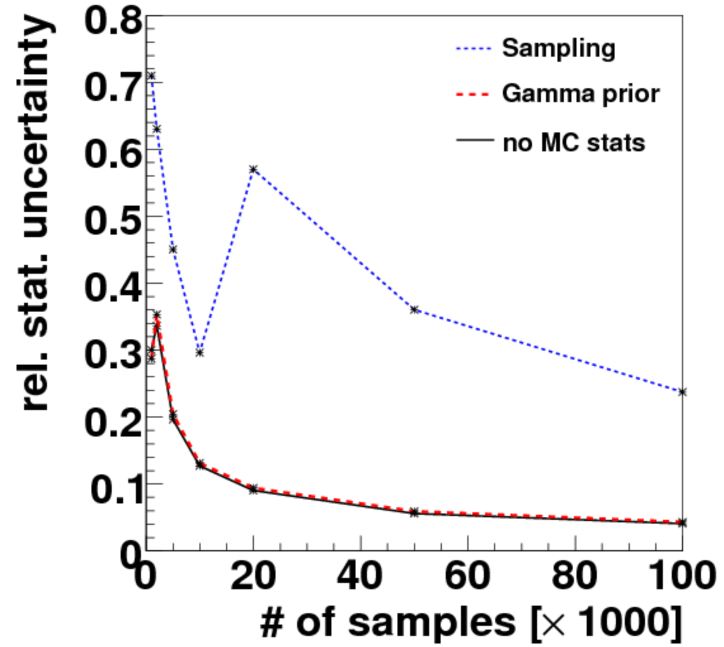
- Systematics posterior is actually sum of individual posteriors from each iSample
- Determine posterior for each systematic sample in limit_bayesian:

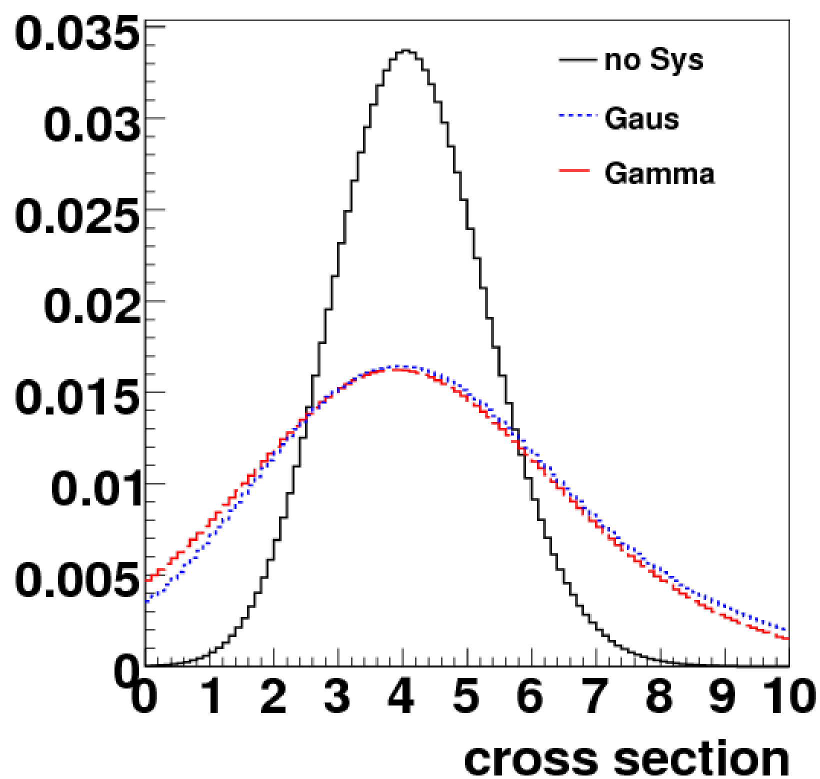
```
for(NSamples systematics samples) {
  for(XS point) {
    val = likelihood_generic(Nobs,sys_bkg[iSample],accL[iSample],XS);
    F[XS] += val;  // later in the code
  }
}
```

- Each of the inputs is an array containing all bins
- Actual code is more complex, has more loops than this, lots of checking of inputs and outputs going on, plus histogram filling
- Plus: First quick evaluation of posterior at only a few points, then full posterior evaluation only for those iSample that have large posterior integral estimate
- Then normalize the posterior sum to unit area later when analyzing posterior

Special systematic: MC statistics uncertainty

- Integration of MC statistics requires large number of samples
- Instead, integrate MC statistics uncertainty analytically
  - Using Gamma prior instead of Gaussian prior
  - Introduces slight bias
    - No problem as long as MC statistics uncertainty is small contribution
- Special sys name: Mcstats
- Integration in poisson_gamma in climit.cpp



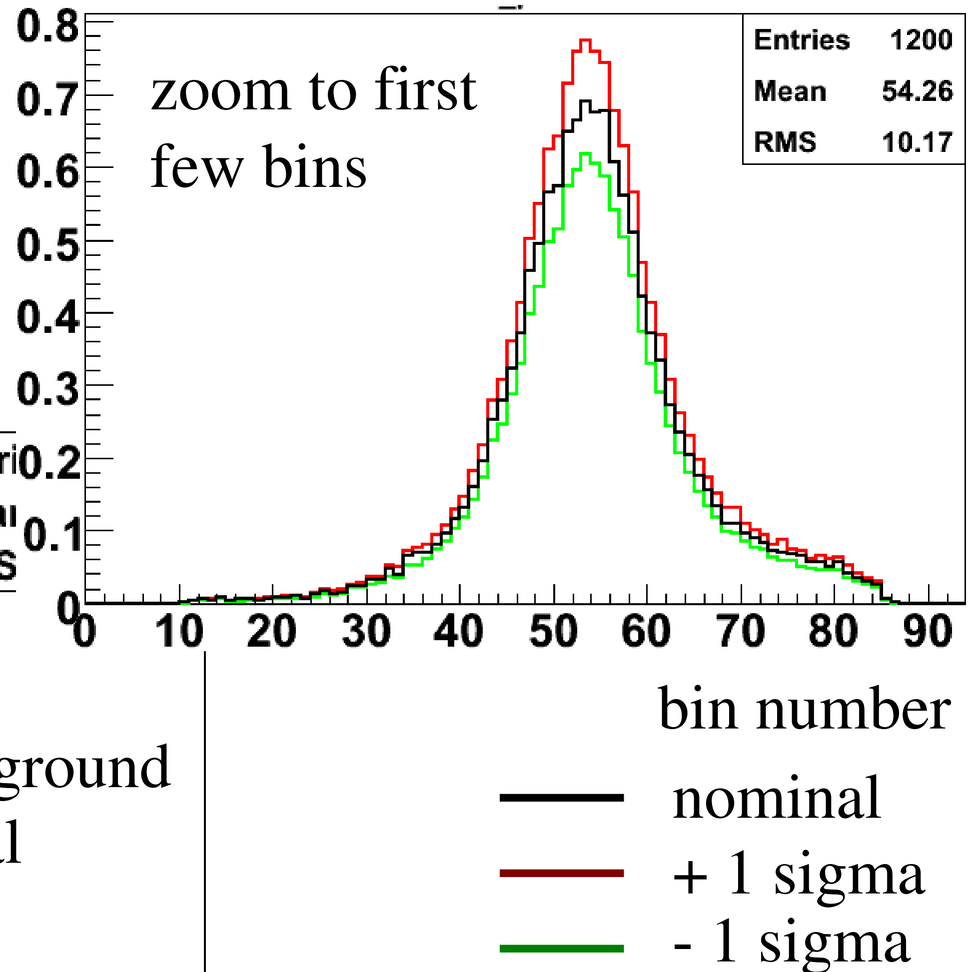


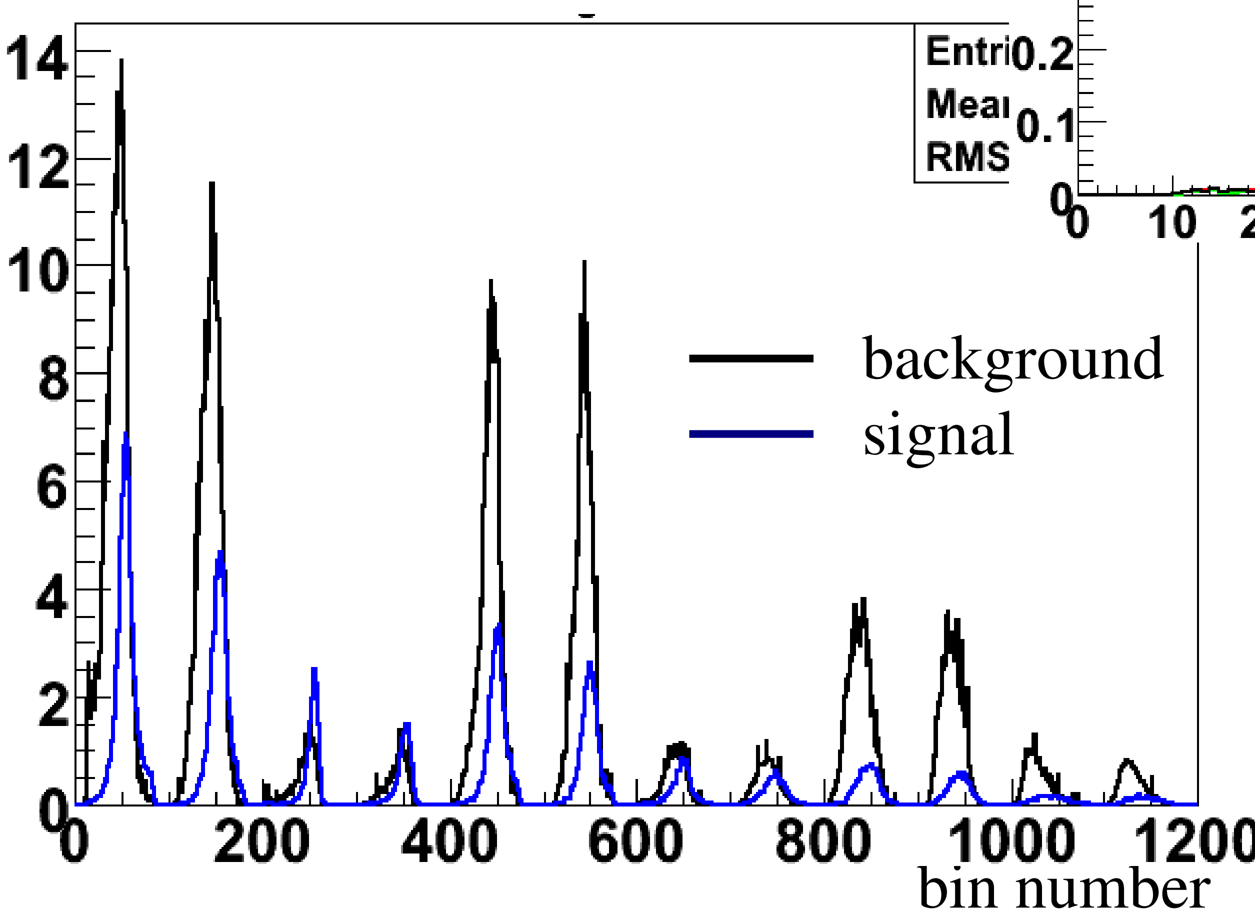
Debug/Info histograms

- Output to screen
  - Program progress
  - Cross section measurement
  - limit
- Histograms and plots in root file
- Background sum and acc*L for all bins as used
  - Including systematic uncertainties, added in quadrature
- Distribution of Gaussian random numbers for each systematic name
- Posterior with peak position and uncertainty
  - Can also do 2d posterior in case of 2 signals
- Systematics posterior

Input distribution

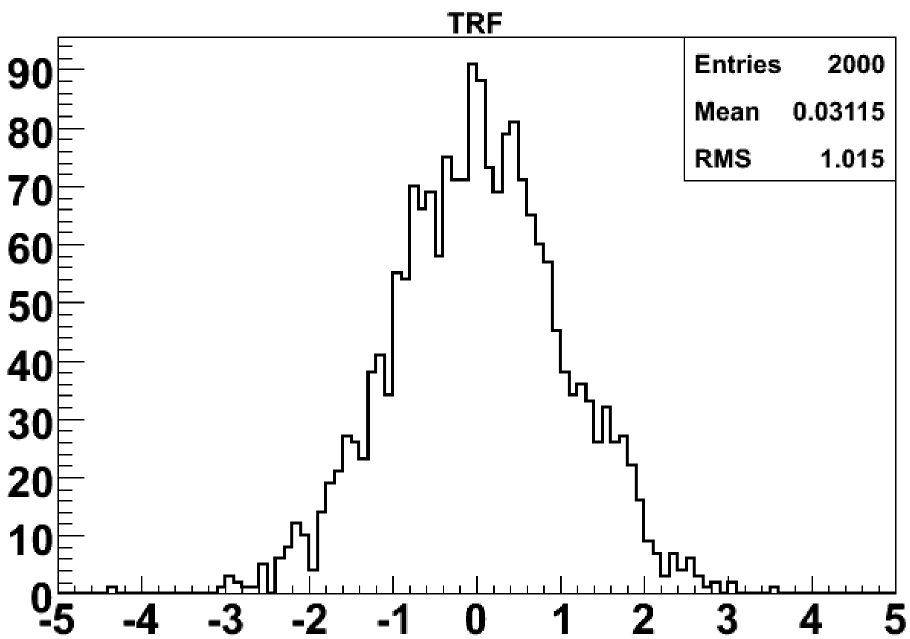
- D0 single top example

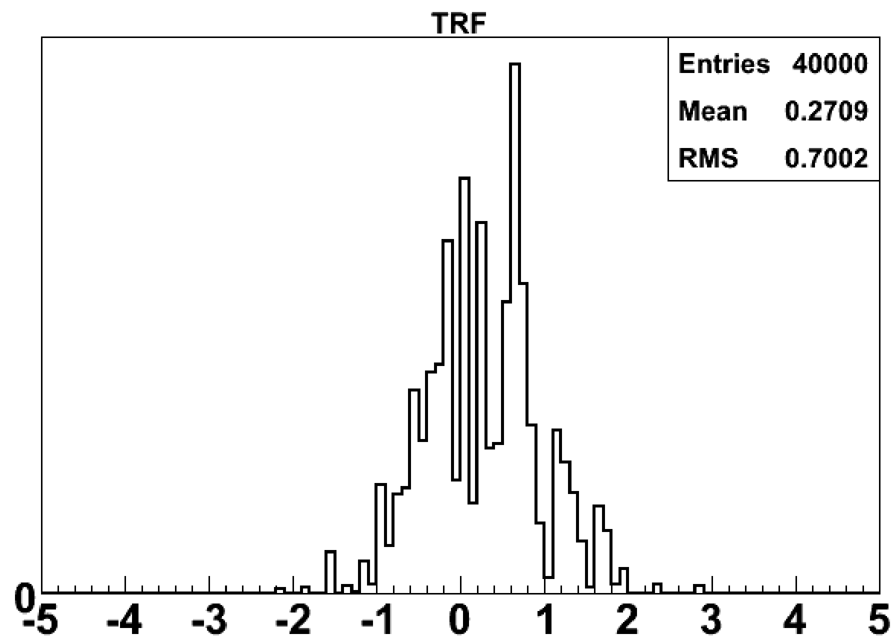




Systematic uncertainties

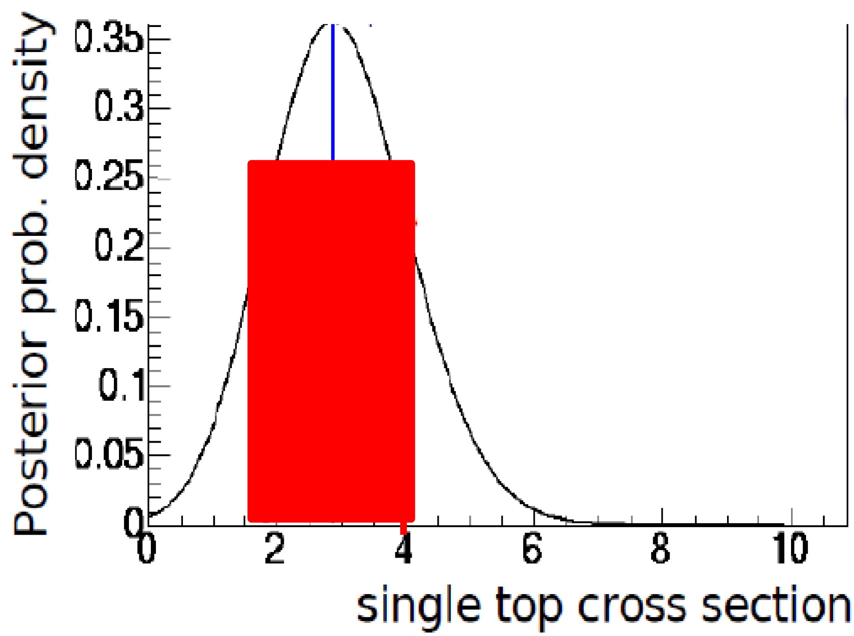
- Integration over systematics is done by sampling from Gaussian distribution, then summing
  - Shown on left
- Rather than summing, histogram this systematic, with posterior weights
  - Systematics histogram, integrated over posterior (right)





Bayes factor, Bayes ratio

- We can get an equivalent of a significance out of the Bayesian posterior
  - Bayes factor: Integral over peak region divided by 0-signal
  - Need to specify in input what area to integrate over
    - Signal.XS
      Signal.XS.Error
- Alternative: Bayes factor
  - Peak height over 0-XS height



- Interpret these as p-value equivalent, then take TMath::NormQuantile(1-p)
- Not widely used, no clear interpretation

Frequentist Analysis

Frequentist statistics

- Only statements about true value, not measured
  - What if I had repeated the experiment many times?
- In top_statistics, done through ensemble testing:
  - How does this actual data experiment compare with ensembles of pseudo-data?
- Ensemble of background-only pseudo-datasets
  - Generate ~∞ # of background-only pseudo-datasets
- Compute log-likelihood ratio for each pseudo-dataset
- Count how many background-only pseudo-datasets have LLR ≥ data
  - Or ≥ mean of a sig+bkg ensemble

Ensemble generation

- Read in sources and bins and channels exactly as for Bayesian limit setting
- Sample from systematic uncertainties
  - Same code/procedure as MC integration
- Then calculate background sum in each channel for this particular set of systematic shifts
- Then draw random Poisson number for this background sum in this channel
  - Or for background+signal if required
- Store bin counts in text file, one line per pseudo-dataset

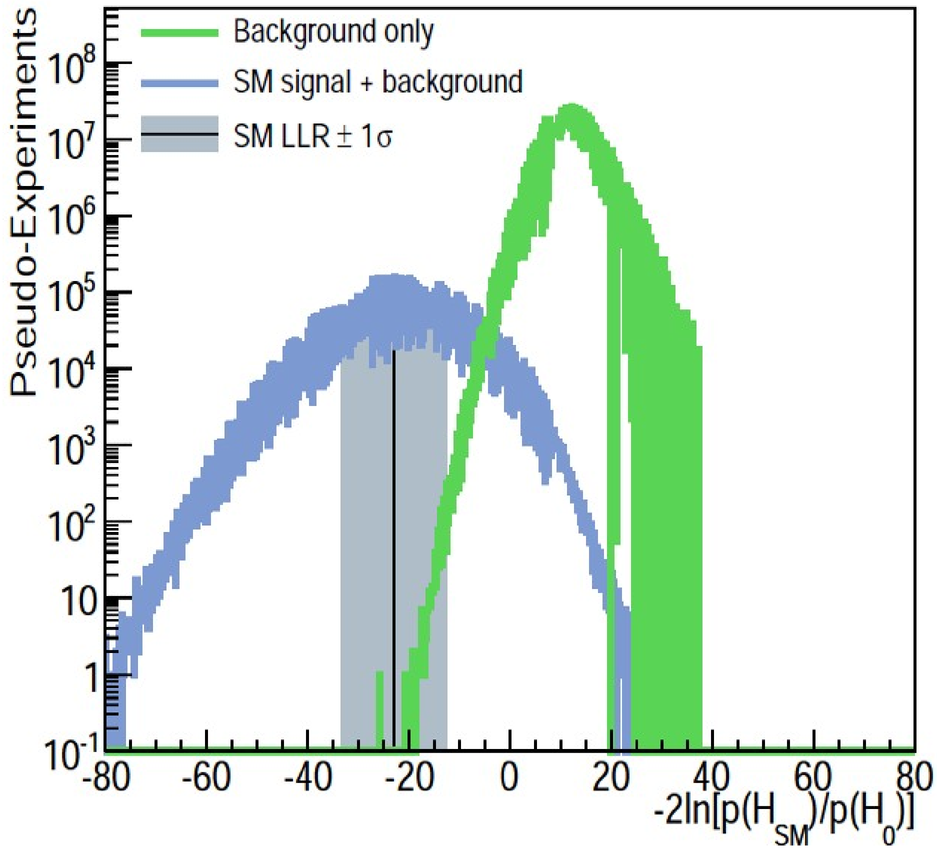
Log-likelihood ratio

- Use Bayesian code to calculate log-likelihood ratio significance
  - LLR, also used in all Tevatron Higgs analyses
- Procedure: generate pseudo-datasets, calculate LLR value for each:
  - Compare null hypothesis (H_0, background only) and alternative hypothesis (H_1 or H_{SM}, signal+background)
    - Compute likelihood of observing background-only $p(H_0)$ and of observing SM signal + background $p(H_{SM})$
    - Likelihood is again just Poisson probability
      $p(H_0) = \text{Poisson}(n_{obs} \mid n_{bkg})$
      $p(H_{SM}) = \text{Poisson}(n_{obs} \mid \text{SM signal} + n_{bkg})$
    - Form test statistic $LLR = -2 \ln[\, p(H_{SM}) / p(H_0) \,]$

LLR in practice

- $LLR = -2 \ln[\, p(H_{SM}) / p(H_0) \,]$
- If no systematics:
  - $p(H_0)$ = Poisson(data | background only)
  - $p(H_{SM})$ = Poisson(data | signal+background)
- With systematics:
  - Integrate over systematics Bayesian style to compute both p's
- Store Poisson values in array to speed up code
  - Need to evaluate LLR for millions of pseudo-dataset
- p-value is fraction of bkg-only pseudo-datasets with LLR value smaller than SM peak
- Convert to Gaussian significance using TMath::NormQuantile(1-p)

LLR distribution



– p-value as probability to observe LLR value seen in data or something more extreme (lower)

top_statistics details

Limit setting code

- Code developed for D0 single top analysis by Harrison Prosper, Supriya Jain, Brigitte Vachon, RS
  - Underlying Bayesian analysis by Harrison Prosper
  - Contributions by Gordon Watts, Dag Gillberg, Aran Garcia-Bellido, Benoit Clement and others
  - Original version developed for first single top analysis in 2004
  - Now also used for Tevatron combination
- Ported to ATLAS by RS

Code structure

- C++ user interface
  - limit_bayesian
- Configuration files using root TEnv
- Underlying Bayesian likelihood calculation in C
  - climit.cpp
- Ensemble generation in ensemblemaker
- Reading in of histograms in limit_base
- Executables for each specific analysis
  - Can do multiple evaluations in one executable
  - Example: ensemble testing: generate, then loop over thousands of pseudo-datasets

Program flow

1) Instantiate limit_bayesian object
   - Set cross section axis, debug flags, Nsamples
2) Read input channels
   - Channel-by-channel
   - For each channel, read list of inputs
     - data, then backgrounds, then signal
     - For each, nominal histogram, then systematics
   - In BDT_helpers.hpp
3) Initialize input distribution
   - Convert channels to long input histogram
   - Generate systematics samples
   - In limit_base
4) Determine Posterior
   - In limit_bayesian, many calls to climit.cpp
5) Analyze posterior (cross section, limit, histograms, ...)

Additional macros

- Posterior plot for publications and talks
  - 1d, 2d, with peak position, uncertainty, limit, etc
  - Vtb evaluation (taking square root of XS)
- BLUE combination
  - Generation of pseudo-datasets correlated between multiple analysis methods
  - Analysis of the resulting cross sections
- LLR plots

Conclusions

- Bayesian and Frequentist statistics both are useful for certain questions
  - All systematic uncertainties are treated in Bayesian fashion
  - Significance well defined using Frequentist statistics
- top_statistics provides statistical analysis tools
  - Bayesian posteriors
  - Tools to analyze them, measure cross sections and set limits
  - Frequentist ensemble testing
  - Macros for pretty plots
- If you need another tool or have a question, let me know!